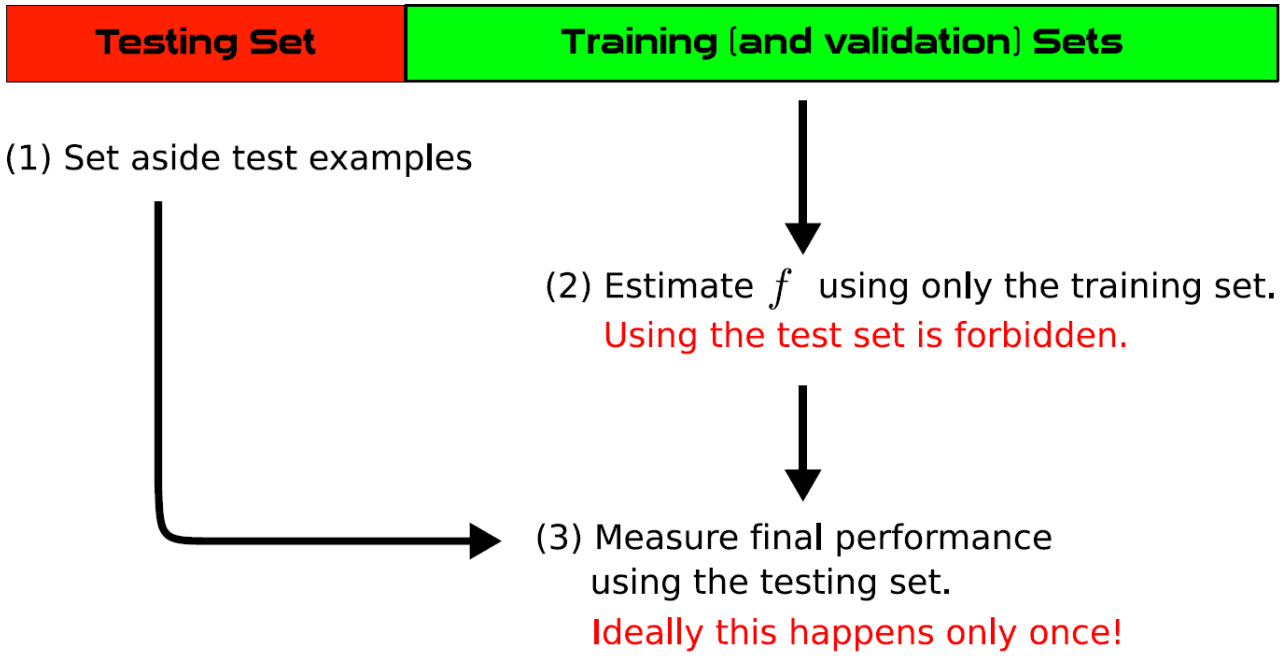
large-scale machine learning revisited

Léon Bottou
Microsoft Research (NYC)

1 three frequent ideas in machine learning.

independent and identically distributed data



- This experimental paradigm has driven machine learning progress.
- The essential assumption is that training and testing data are exchangeable, e.g., follow the same distribution.

model selection tradeoffs

Approximation

- We cannot search f^* among all possible functions.
- We search instead $f^*_{\mathcal{F}}$ that minimizes the expected risk $E(f)$ within some richly parameterized family of functions $\mathcal{F}$.

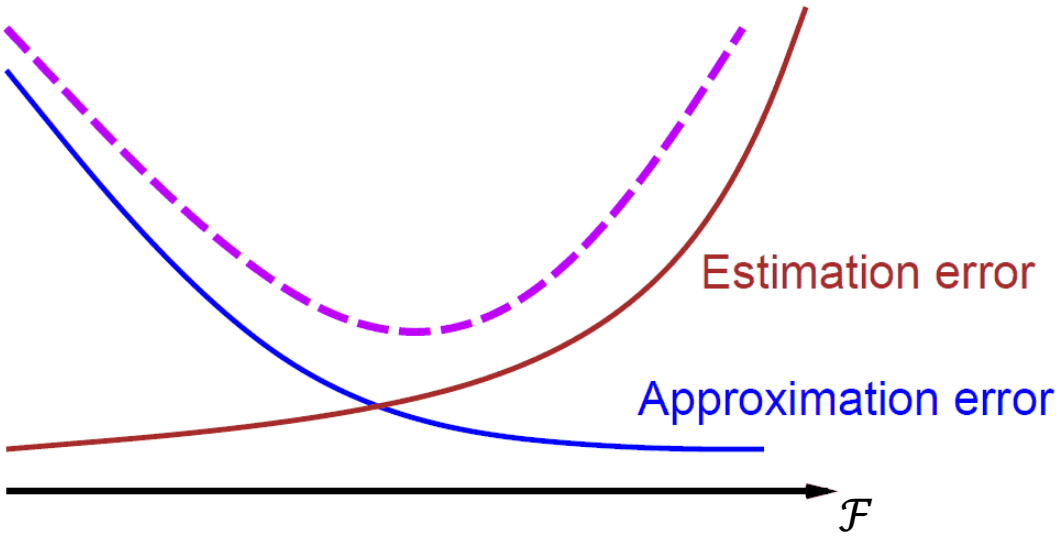
Estimation

- We cannot minimize $E(f)$ because the data distribution is unknown.
- We minimize instead the empirical risk $E_n(f)$

$$E_n(f) = \frac{1}{n} \sum \ell(f(x_i), y_i)$$

model selection tradeoffs

$$E(f_n) - E(f^*) = \left(E(f_F^*) - E(f^*)\right) \quad \text{Approximation Error}$$
$$+ \left(E(f_n) - E(f_{\mathcal{F}}^*)\right) \quad \text{Estimation Error}$$



How complex a model can you afford with your data?

Vapnik's razor

" When solving a (learning) problem of interest, do not solve a more complex problem as an intermediate step. "

How complex a model can you afford with your data? (again)

- To classify patterns, use a model that outputs a class and nothing else.

- To achieve something more complex,
  i. carefully define the problem,
  ii. solve the problem and nothing else.

conceptual viewpoints in machine learning

	capacity tradeoff	same distribution	Vapnik's razor
statistical learning theory (ERM, SRM, SVM, …)	yes	yes	yes
Bayesian learning (generative models, priors, …)	yes	yes	no [1]
algorithmic learning theory (regret bounds, …)	yes	no [2]	yes

1) See the discriminant versus generative debate. On the one hand, some authors see generative models as an implicit form of regularization. On the other hand, generative models are often appealing because they easy to combine, unlike strict discriminant classifiers.

2) Online regret bounds express how a learning algorithm performs relative to a class of competitors. Although they do not depend on i.i.d. assumptions, they lose value when none of the competitors works well, for instance because the data is too far from i.i.d..

2 the tradeoffs of large-scale learning.

statistics and computation

Statistical Perspective

- It is good to optimize an objective function that ensures a fast estimation rate when the number of examples increases.

Optimization Perspective

- To efficiently solve large problems, it is preferable to choose an optimization algorithm with strong convergence properties.

Incorrect Conclusion

- To address large-scale learning problems, use the best algorithm to optimize an objective function with fast estimation rates. *

statistics and computation

- Baseline large-scale learning algorithm

Randomly discarding data is the simplest way to handle large datasets.

  - What is the statistical benefit of processing more data?
  - What is the computational cost of processing more data?

- We need a theory that links Statistics and Computation!

- 1967: Vapnik's theory does not discuss computation.
- 1981: Valiant's learnability excludes exponential time algorithms, but (i) polynomial time already too slow, (ii) few actual results.

learning with approximate optimization

Computing $f_n = \arg\min_{f \in \mathcal{F}} E_n(f)$ is often costly.

Since we already optimize a **surrogate** function
why should we compute its optimum f_n exactly?

Let's assume our optimizer returns $\tilde{f}_n$
such that $E_n(\tilde{f}_n) < E_n(f_n) + \rho$.

For instance, one could stop an iterative
optimization algorithm long before its convergence.

error decomposition

$$
\begin{aligned}
E(\tilde{f}_n) - E(f^*) &= E(f^*_{\mathcal{F}}) - E(f^*) && \text{Approximation error} \\
&+ E(f_n) - E(f^*_{\mathcal{F}}) && \text{Estimation error} \\
&+ E(\tilde{f}_n) - E(f_n) && \text{Optimization error}
\end{aligned}
$$

Problem:

Choose $\mathcal{F}$, n, and ρ to make this as small as possible,

subject to budget constraints $\begin{cases} \text{max number of examples } n \\ \text{max computing time } T \end{cases}$

small scale versus large scale

Beyond informal definitions…

Small scale learning problem

- We have a small-scale learning problem when the active budget constraint is the number of examples n.

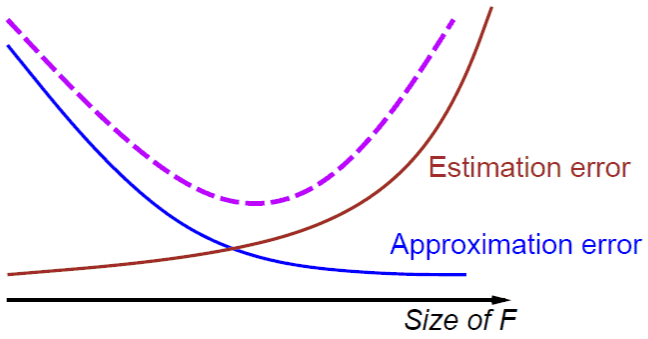
Large-scale learning problem

- We have a large-scale learning problem when the active budget constraint is the computing time T.

small-scale learning

The active budget constraint is the number of examples.

- To reduce the estimation error, take n as large as the budget allows.
- To reduce the optimization error to zero, take $\rho = 0$.
- We need to adjust the size of $\mathcal{F}$.



See Structural Risk Minimization (Vapnik 74) and later works.

large-scale learning

The active budget constraint is the computing time.

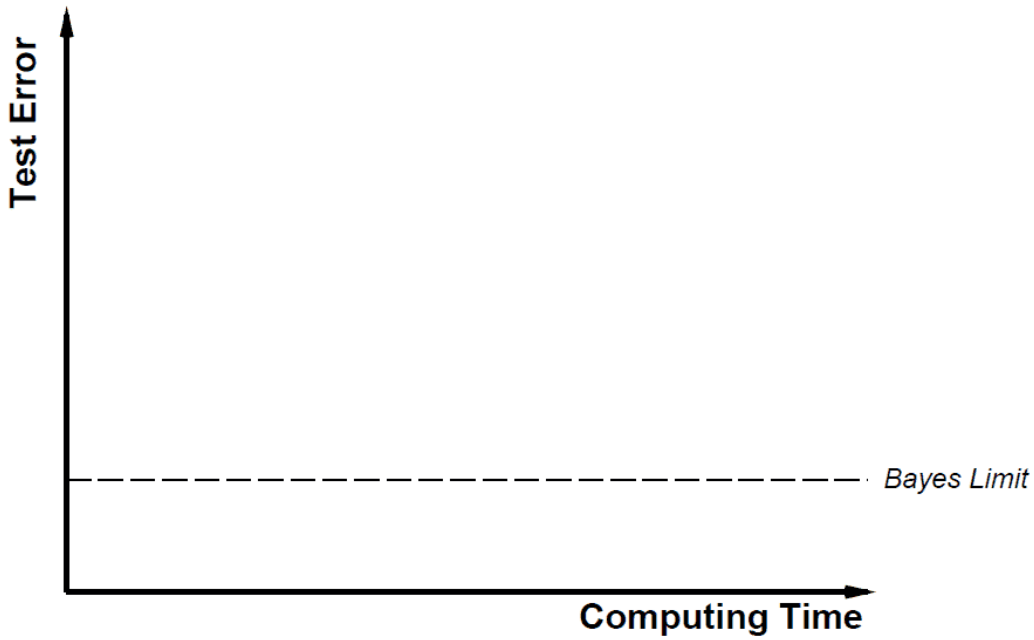
- More complicated tradeoffs.
  The computing time depends on the three variables: $\mathcal{F}$, n, and ρ.

- Example.
  If we choose ρ small, we decrease the optimization error. But we must also decrease $\mathcal{F}$ and/or n with adverse effects on the estimation and approximation errors.

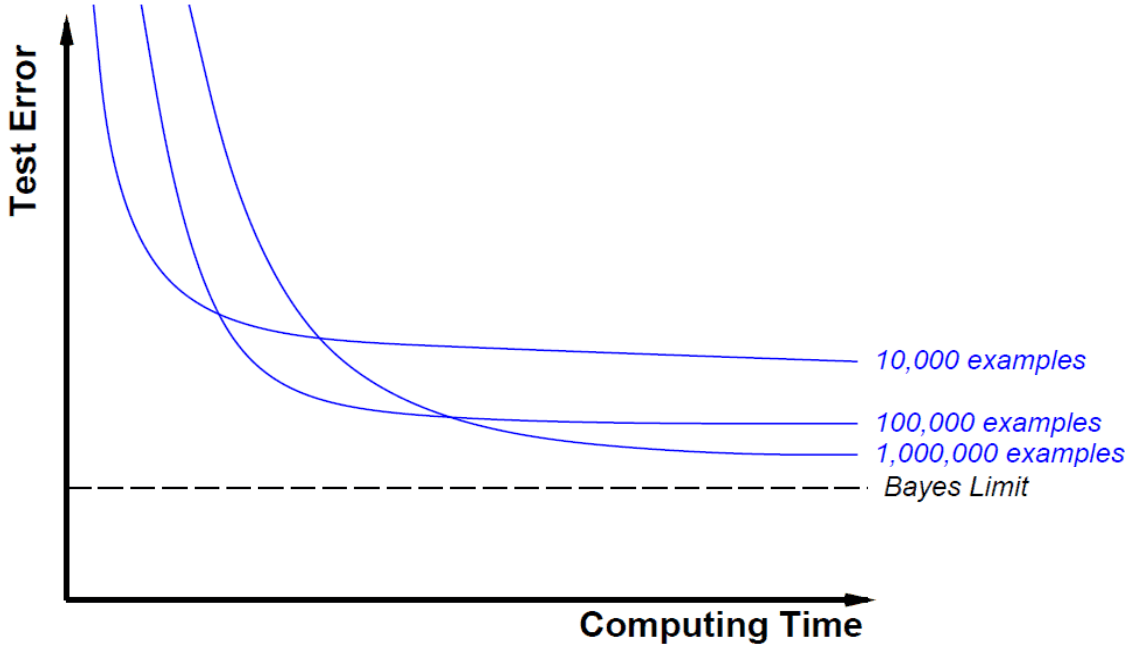
- The exact tradeoff depends on the optimization algorithm.

- We can compare optimization algorithms rigorously.

test error versus training time

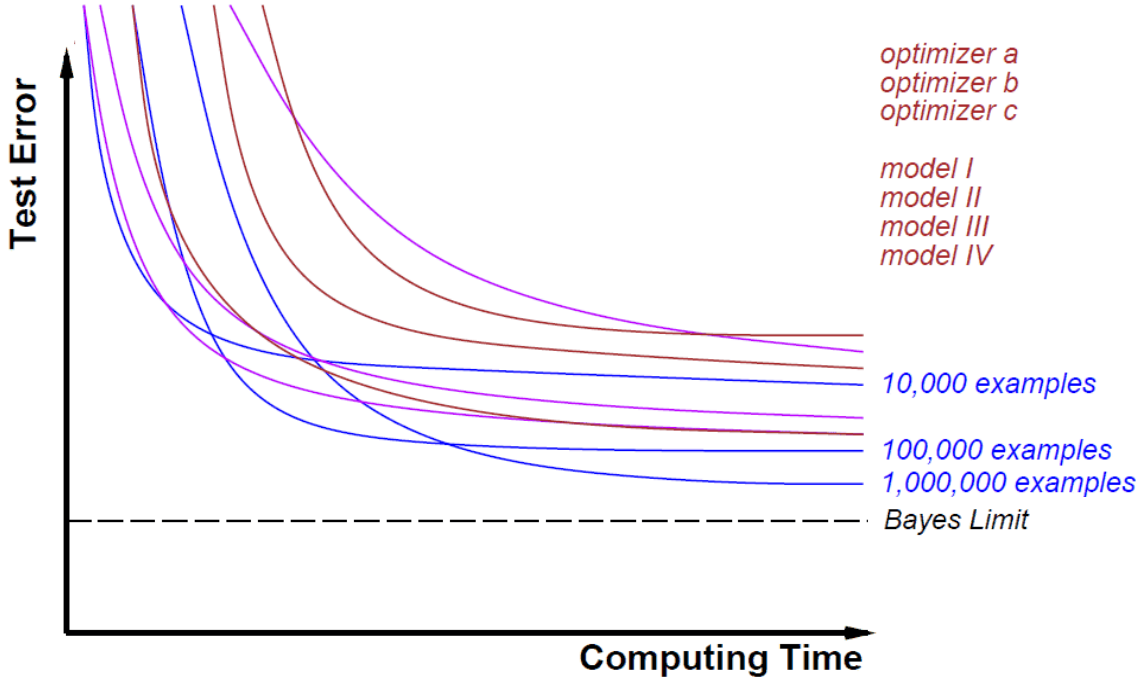


test error versus training time



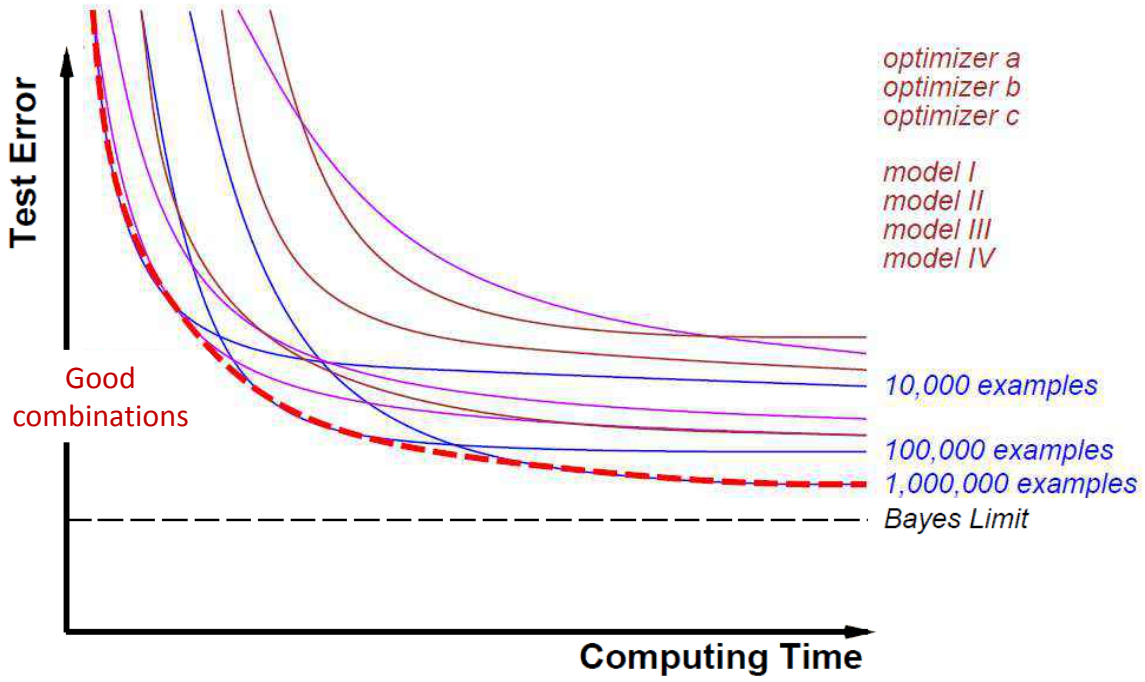
- Vary the number of examples

test error versus training time



- Vary the number of examples, the model, the algorithm

test error versus training time



- Optimal combination depends on training time budget.

asymptotics

$$
\begin{aligned}
E(\tilde{f}_n) - E(f^*) &= E(f^*_{\mathcal{F}}) - E(f^*) && \text{Approximation error} \\
&+ E(f_n) - E(f^*_{\mathcal{F}}) && \text{Estimation error} \\
&+ E(\tilde{f}_n) - E(f_n) && \text{Optimization error}
\end{aligned}
$$

Asymoptotic Approach

All three errors must decrease with comparable rates.

Forcing one of the errors to decrease much faster

- would require additional computing efforts,
- but would not significantly improve the test error.

asymptotics: estimation

Uniform convergence bounds

$$\text{Estimation error} \leq \mathcal{O}\left(\left[\frac{d}{n}\log\frac{n}{d}\right]^{\alpha}\right) \text{ with } \frac{1}{2} \leq \alpha \leq 1 \ .$$

Value d describes the *capacity* of our system.

The simplest capacity measure is the *Vapnik-Chervonenkis* dimension of $\mathcal{F}$.

There are in fact three (four?) types of bounds to consider:

- Classical V-C bounds (pessimistic): $\mathcal{O}\left(\sqrt{\frac{d}{n}}\right)$
- Relative V-C bounds in the realizable case: $\mathcal{O}\left(\frac{d}{n}\log\frac{n}{d}\right)$
- Localized bounds (variance, Tsybakov): $\mathcal{O}\left(\left[\frac{d}{n}\log\frac{n}{d}\right]^{\alpha}\right)$

Fast estimation rates: (Bousquet, 2002; Tsybakov, 2004; Bartlett et al., 2005; . . .)

asymptotics: estimation + optimization

Uniform convergence arguments give

$$\text{Estimation error} + \text{Optimization error} \leq \mathcal{O}\left(\left[\frac{d}{n}\log\frac{n}{d}\right]^{\alpha} + \rho\right) = \varepsilon \ .$$

This is true for all three cases of uniform convergence bounds.

➡ **Scaling laws for ρ when $\mathcal{F}$ is fixed**

The approximation error is constant.

- No need to choose ρ smaller than $\mathcal{O}\left(\left[\frac{d}{n}\log\frac{n}{d}\right]^{\alpha}\right)$.
- Not advisable to choose ρ larger than $\mathcal{O}\left(\left[\frac{d}{n}\log\frac{n}{d}\right]^{\alpha}\right)$.

...estimation + optimization + approximation

When $\mathcal{F}$ is chosen via a λ-regularized cost

- Uniform convergence theory provides bounds for simple cases
  (Massart-2000; Zhang 2005; Steinwart et al., 2004-2007; ...)
- Scaling laws for n, λ and ρ depend on the optimization algorithm.
- 

See (Shalev-Shwartz and Srebro, ICML 2008) for Linear SVMs.

When $\mathcal{F}$ is realistically complicated

Large datasets matter

- because one can use more features,
- because one can use richer models.

Bounds for such cases are rarely realistic enough.

analysis of a simple case

Simple parametric setup

- Family of function $\mathcal{F}$ fixed.
- Functions $f_w(x)$ are linearly parametrized by $w \in \mathbb{R}^d$.

Comparing three iterative optimization algorithms

1. Gradient descent
2. Second order gradient descent (Newton)
3. Stochastic gradient descent

quantities of interest

- Empirical Hessian at the empirical optimum w_n.

$$H = \frac{\partial^2 E_n}{\partial w^2}(f_{w_n}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ell(f_n(x_i), y_i)}{\partial w^2}$$

- Empirical Fisher Information matrix at the empirical optimum w_n.

$$G = \frac{1}{n}\sum_{i=1}^{n} \left[\left(\frac{\partial \ell(f_n(x_i), y_i)}{\partial w} \right) \left(\frac{\partial \ell(f_n(x_i), y_i)}{\partial w} \right)' \right]$$

- **Condition number**

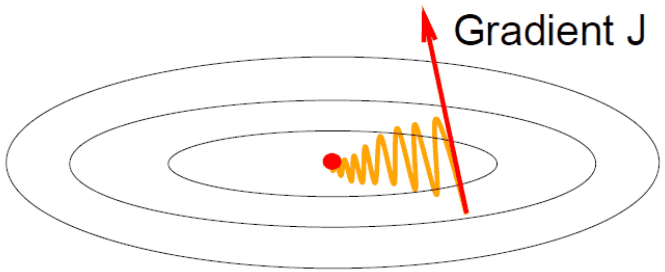
  We assume that there are $\lambda_{\min}$, $\lambda_{\max}$ and ν such that
  - $\operatorname{trace}(GH^{-1}) \approx \nu$.
  - $\operatorname{spectrum}(H) \subset [\lambda_{\min}, \lambda_{\max}]$.

  and we define the condition number $\kappa = \lambda_{\max}/\lambda_{\min}$.

gradient descent

Iterate

- $w_{t+1} \leftarrow w_t - \eta \frac{\partial E_n(f_{w_t})}{\partial w}$



Best speed achieved with fixed learning rate $\eta = 1/\lambda_{\max}$.
(e.g., Dennis & Schnabel, 1983)

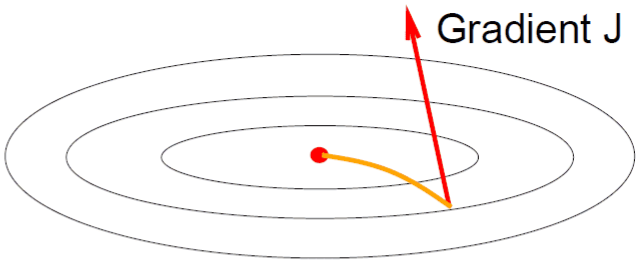
	Cost per iteration	Iterations to reach ρ	Time to reach accuracy ρ	Time to reach $E(\tilde{f}_n) - E(f^*_{\mathcal{F}}) < \varepsilon$
GD	$\mathcal{O}(nd)$	$\mathcal{O}\left(\kappa \log \frac{1}{\rho}\right)$	$\mathcal{O}\left(nd\kappa \log \frac{1}{\rho}\right)$	$\mathcal{O}\left(\frac{d^2\kappa}{\varepsilon^{1/\alpha}} \log^2 \frac{1}{\varepsilon}\right)$

– In the last column, n and ρ are chosen to reach ε as fast as possible.
– Solve for ε to find the best error rate achievable in a given time.
– Remark: abuses of the $\mathcal{O}()$ notation

second order gradient descent

Iterate

- $w_{t+1} \leftarrow w_t - H^{-1} \frac{\partial E_n(f_{w_t})}{\partial w}$



We assume H^{-1} is known in advance.
Superlinear optimization speed (e.g., Dennis & Schnabel, 1983)

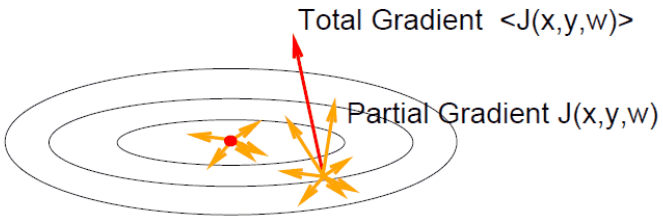
	Cost per iteration	Iterations to reach ρ	Time to reach accuracy ρ	Time to reach $E(\tilde{f}_n) - E(f^*_{\mathcal{F}}) < \varepsilon$
2GD	$\mathcal{O}(d(d+n))$	$\mathcal{O}\left(\log\log\frac{1}{\rho}\right)$	$\mathcal{O}\left(d(d+n)\log\log\frac{1}{\rho}\right)$	$\mathcal{O}\left(\frac{d^2}{\varepsilon^{1/\alpha}}\log\frac{1}{\varepsilon}\log\log\frac{1}{\varepsilon}\right)$

– Optimization speed is much faster.
– Learning speed only saves the condition number κ.

stochastic gradient descent

Iterate (Robbins-Monro)

- Draw random example (x_t, y_t).
- $w_{t+1} \leftarrow w_t - \frac{\eta}{t} \frac{\partial \ell(f_{w_t}(x_t), y_t)}{\partial w}$



Best decreasing gain schedule with $\eta = 1/\lambda_{\min}$.
(see e.g. Murata, 1998; Bottou & LeCun, 2004)

	Cost per iteration	Iterations to reach ρ	Time to reach accuracy ρ	Time to reach $E(\tilde{f}_n) - E(f^*_{\mathcal{F}}) < \varepsilon$
SGD	$\mathcal{O}(d)$	$\frac{\nu k}{\rho} + o\left(\frac{1}{\rho}\right)$	$\mathcal{O}\left(\frac{d \nu k}{\rho}\right)$	$\mathcal{O}\left(\frac{d \nu k}{\varepsilon}\right)$

With $1 \leq k \leq \kappa^2$

– Optimization speed is *catastrophic*.
– Learning speed does not depend on the statistical estimation rate α.
– Learning speed depends on condition number κ but *scales very well*.

benchmarking sgd on simple problems

- The theory suggests that SGD is very competitive.
  - Many people associate SGD with trouble.
- SGD historically associated with back-propagation.
  - Multilayer networks are very hard problems (nonlinear, nonconvex)
  - What is difficult, SGD or MLP?

- **Try PLAIN SGD on a simple learning problem.**

Download from `http://leon.bottou.org/projects/sgd`.
These simple programs are very short.

text categorization with a linear svm

- **Dataset**
  - Reuters RCV1 document corpus.
  - 781,265 training examples, 23,149 testing examples.
  - 47,152 TF-IDF features.
- **Task**
  - Recognizing documents of category `CCAT`.
  - Minimize $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\lambda}{2}w^2 + \ell(w\,x_i + b,\, y_i)\right)$.
  - Update $w \leftarrow w - \eta_t \nabla(w_t, x_t, y_t) = w - \eta_t\left(\lambda w + \frac{\partial \ell(w\,x_t + b,\, y_t)}{\partial w}\right)$

Same setup as (Shalev-Schwartz et al., 2007) but plain SGD.

text categorization with a linear svm

- **Results: Linear SVM**

  $\ell(\hat{y}, y) = \max\{0, 1 - y\hat{y}\} \quad \lambda = 0.0001$

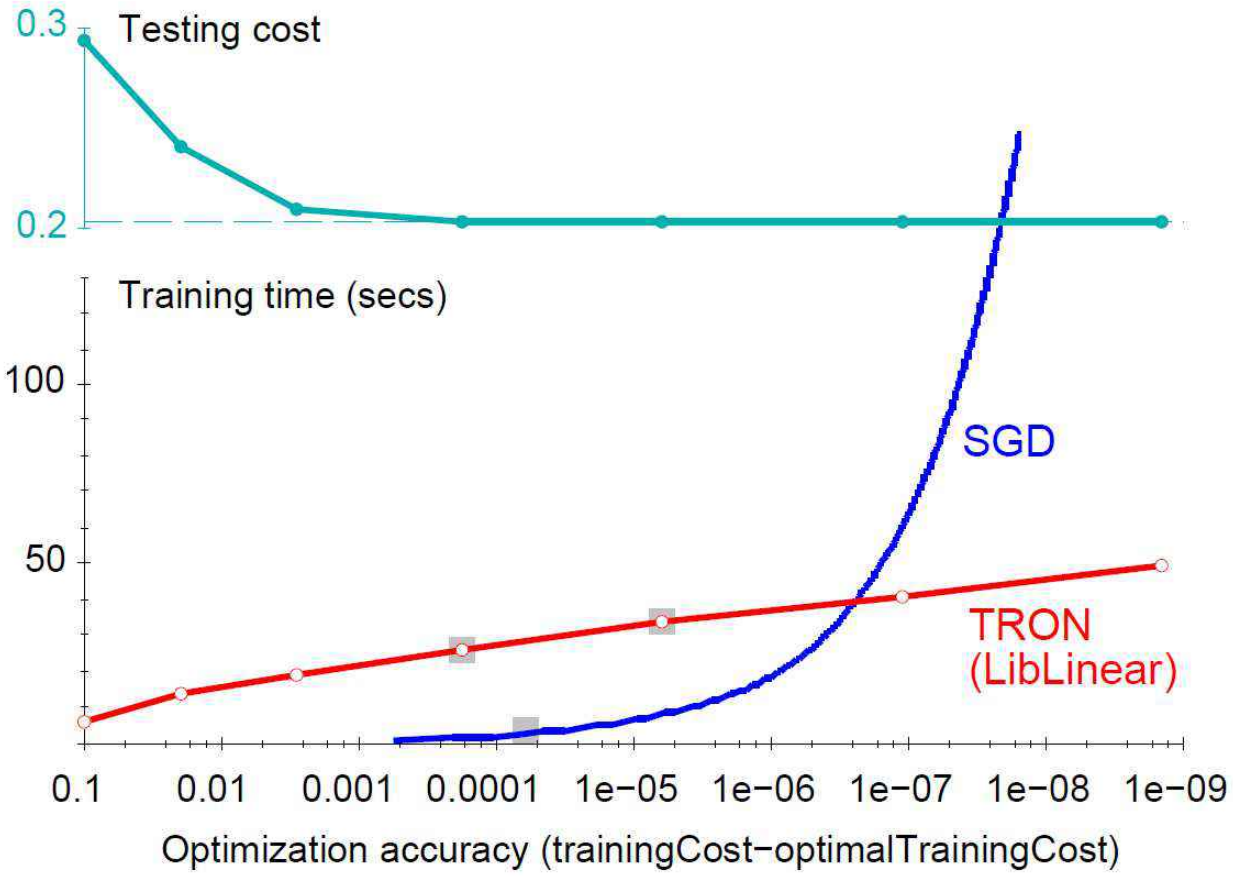
	Training Time	Primal cost	Test Error
SVMLight	23,642 secs	*0.2275*	*6.02%*
SVMPerf	66 secs	0.2278	6.03%
SGD	1.4 secs	*0.2275*	*6.02%*

- **Results: Log-Loss Classifier**

  $\ell(\hat{y}, y) = log(1 + exp(-y\hat{y})) \quad \lambda = 0.00001$

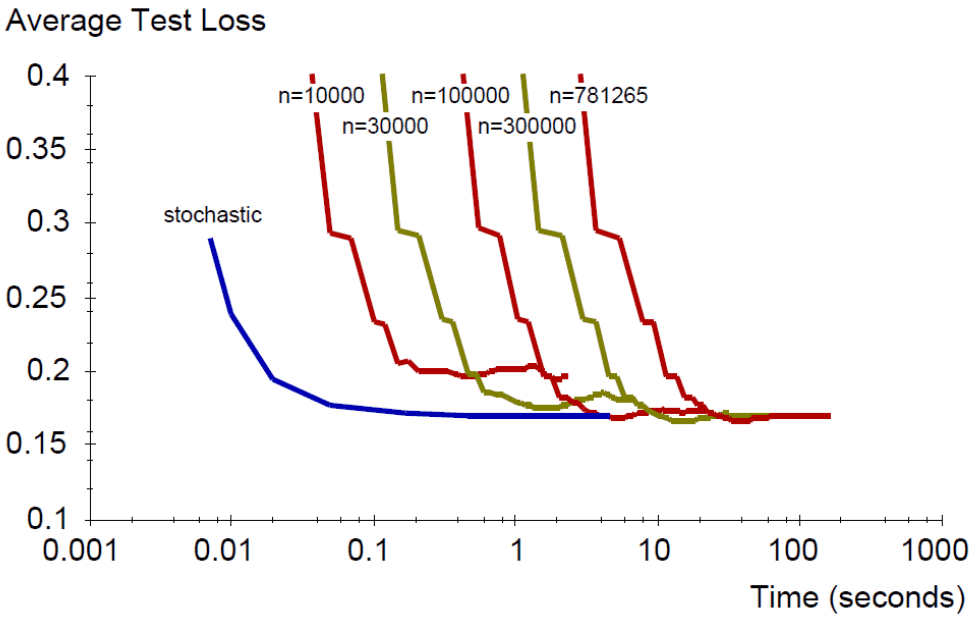
	Training Time	Primal cost	Test Error
TRON(LibLinear, $\varepsilon = 0.01$)	30 secs	0.18907	*5.68%*
TRON(LibLinear, $\varepsilon = 0.001$)	44 secs	*0.18890*	5.70%
SGD	2.3 secs	0.18893	*5.66%*

text categorization with a linear svm



text categorization with a linear svm

From: Olivier Chapelle
Date: Sunday 2007-10-28 22:26:44
. . . you should really run batch with various training set sizes . . .



Log-loss problem

Batch Conjugate Gradient on various training set sizes

Stochastic Gradient on the full set

Why is SGD near the enveloppe?

text chunking with a crf

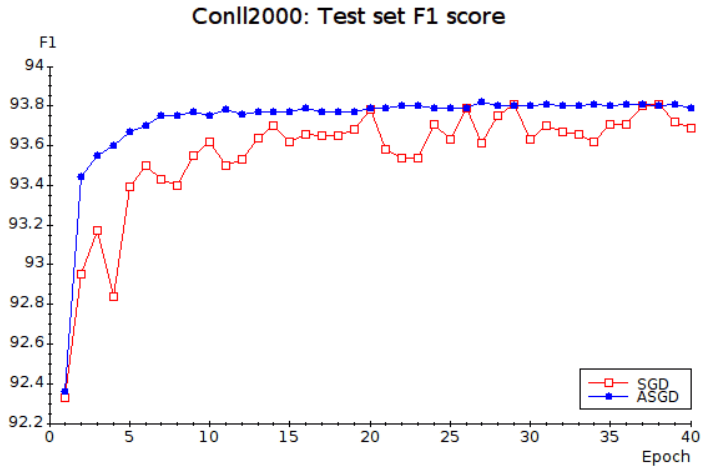
- **Dataset**
  - CONLL 2000 Chunking Task:
    Segment sentences in syntactically correlated chunks
    (e.g., noun phrases, verb phrases.)
  - 106,978 training segments in 8936 sentences.
  - 23,852 testing segments in 2012 sentences.
- **Model**
  - Conditional Random Field (all linear, log-loss.)
  - Features are n-grams of words and part-of-speech tags.
  - 1,679,700 parameters.

Same setup as (Vishwanathan et al., 2006) but plain SGD.

text chunking with a crf

- **Results**

	Training Time	Primal cost	Test F1 score
L-BFGS	4335 secs	*9042*	93.74%
SGD	568 secs	9098	*93.75%*
ASGD	*135* secs	9325	*93.79%*



the tradeoffs of large-scale learning

Small-scale learning ≠ large-scale learning

- Large-scale learning involves more complex tradeoffs
  that depends on the properties of the optimization algorithm.

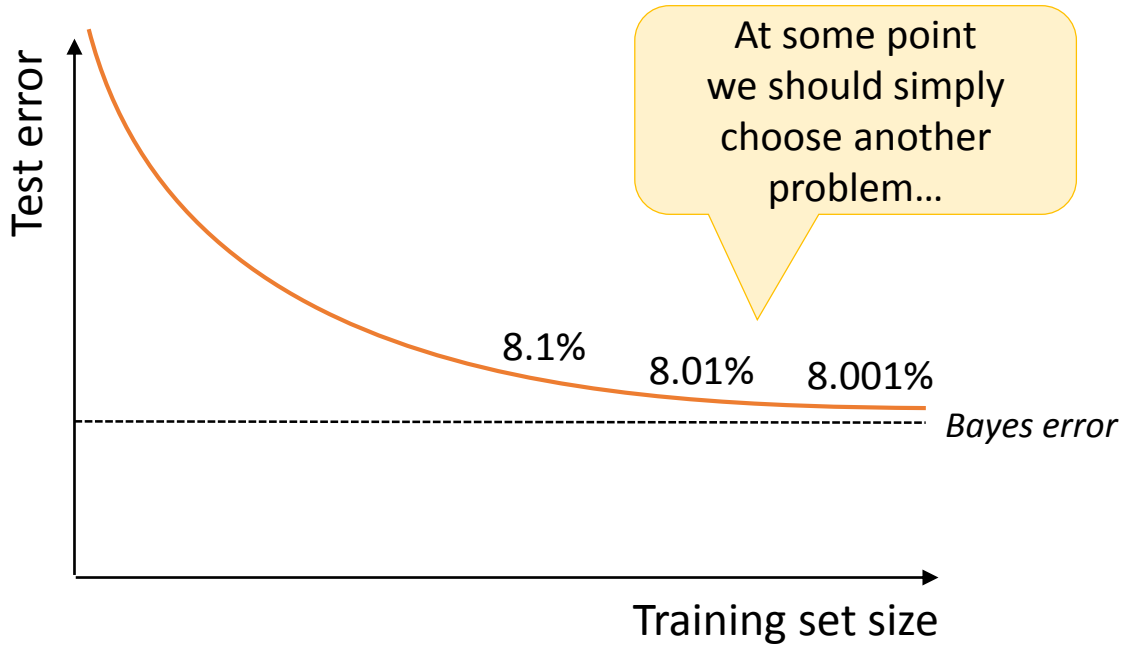
Good optimization algorithm ≠ good learning algorithm

- Mediocre optimization algorithms (e.g., SGD)
  often outperform sophisticated optimization algorithms
  on large-scale learning problems.

provided that the code is correct (which is harder than it seems.)

3- breadth versus accuracy

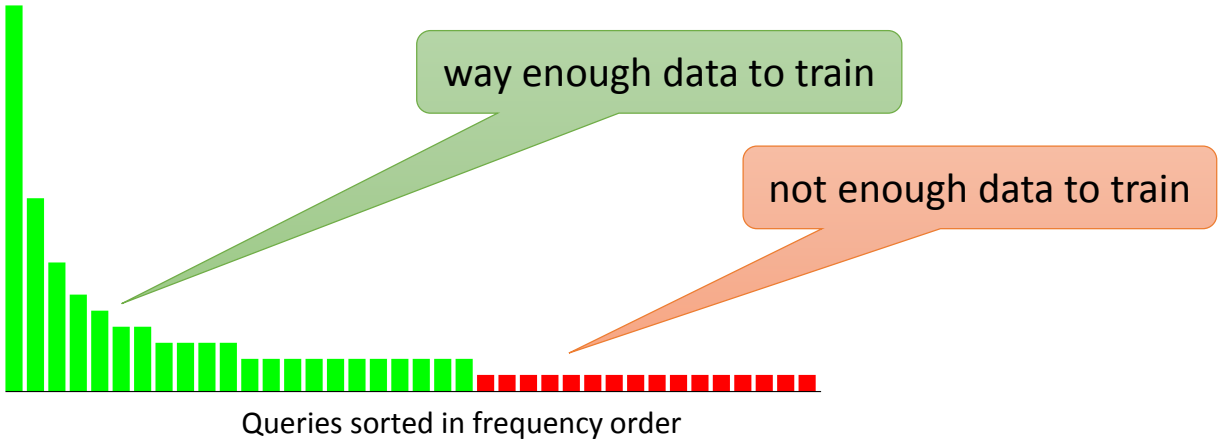
diminishing returns



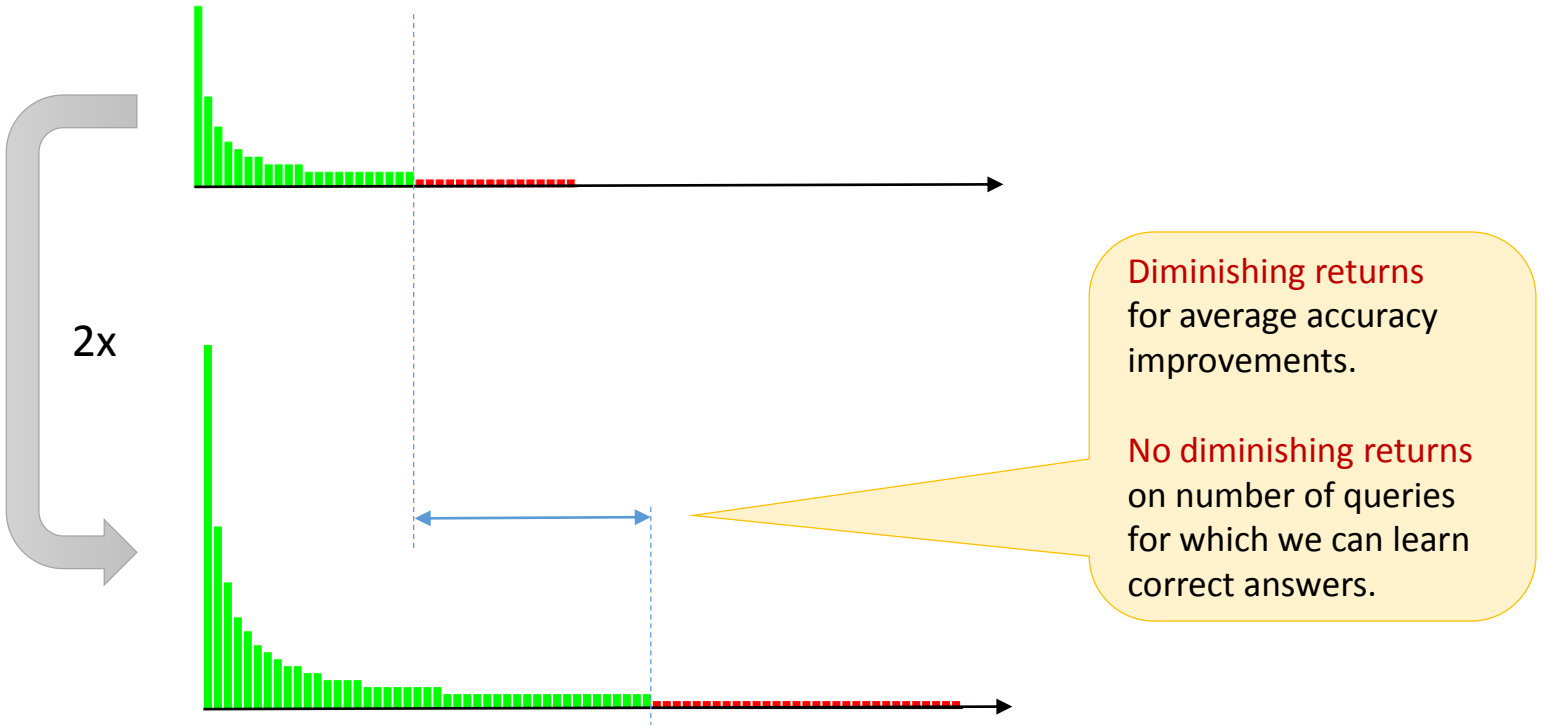
- Accuracy improvements cannot justify the computational cost forever.
- Why then use very large training sets ?

Zipf distributed data

- Roughly half of the search queries are unique.



doubling the size of the training set

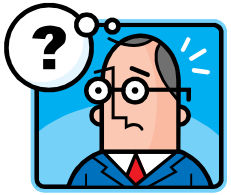


the value of big data in machine learning

Accuracy improvements are subject to diminishing returns.
Breadth improvements are not subject to diminishing returns.

"How accurately do we recognize an object category?"
vs. "How many categories do we recognize well enough?"

Should we optimize a different criterion?



How does this helps if average accuracy is what we care about ?

same distribution?

Data collection in traditional machine learning

- Training data collection for real-life machine learning is difficult.
  The data distribution must reflect the operational conditions.
- The i.i.d. assumption is not automatically satisfied.
  It happens through manual data curation.

Data collection in big data machine learning

- Big data exists because data collection is automated.
- No manual curation to enforce the identical distribution assumption.
- The output of the machine learning system frequently
  impacts the distribution of future training data.

dealing with covariate shifts

$$P(X,Y) = P(Y|X)\ P(X)$$

We want to model $Y \approx f(X)$.
We must assume that the training data describes $P(Y|X)$ well enough.

We cannot trust $P(X)$.
We want to train a system robust to $P(X)$ changes.

Minimizing the training set error

- Approximation errors are pushed towards patterns X with low probability.
- What if these patterns occur more frequently at testing time?

Maximize the "diversity" of patterns that are recognized well enough.

- Yields a solution that is more robust to $P(X)$ changes.

independent examples?

MNIST digits

- Training set : 600 writers × 100 digits.
- Testing set: 100 different writers × 100 digits.
- How many independent training examples, 600 or 60000?

Search queries

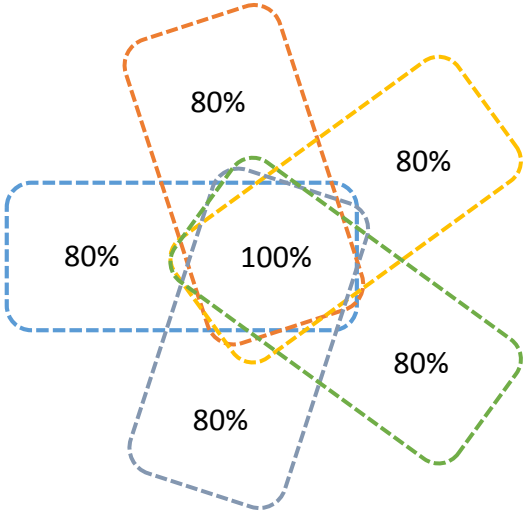
bing wrong trousers

- Lots of queries entered by lots of users.
- The search engines must satisfy the users, not the queries.
- The satisfaction of a user is not proportional to its satisfied queries.

independent examples?

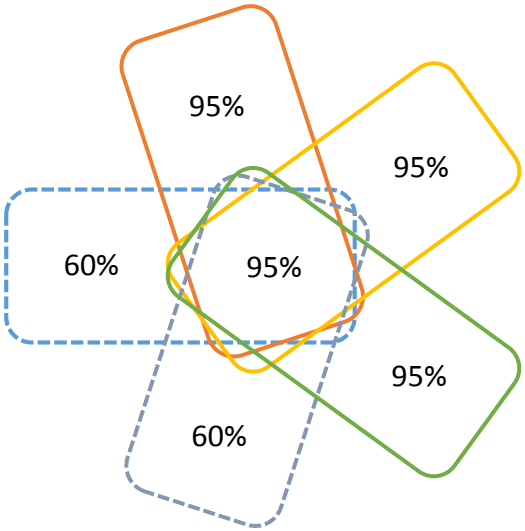
Assume that a user is not satisfied below 95% correct answers.

Minimize average error



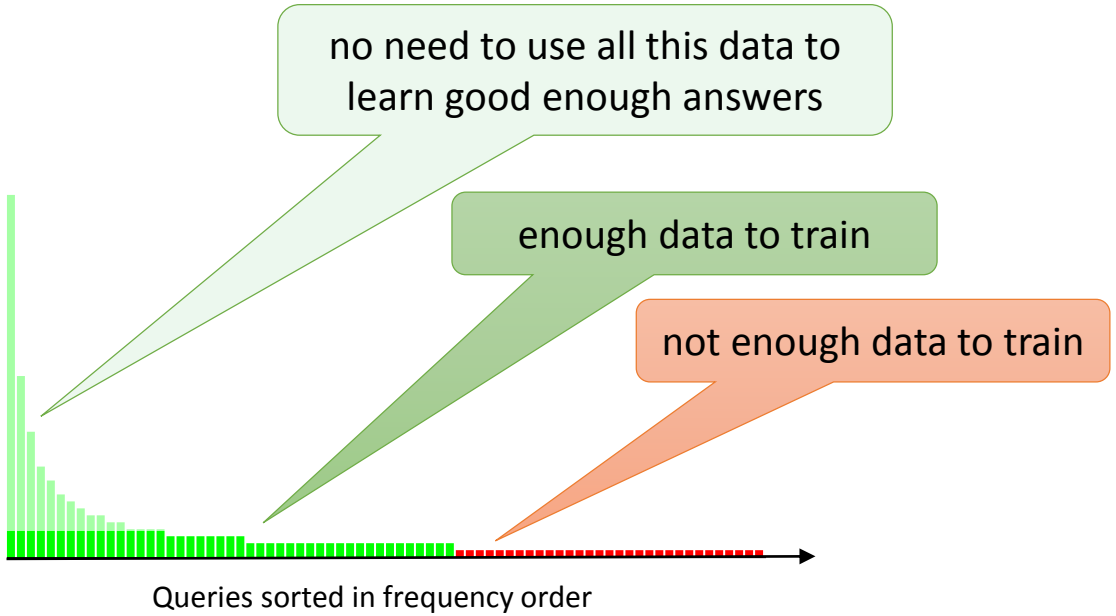
Average error: 10%
Users satisfied: 0

Maximize queries answered at 95% level



Average error: 12%
Users satisfied: 3

scalability opportunities



- No need to consider all examples of already known queries.
- Best is to focus on queries near the boundary of the known area.
- Curriculum learning and active learning come naturally in this context.
- Scalability gains across the board.

3. deep learning and transfer learning

engineering machine learning

Reading check amounts

– Input $\mathcal{X}$: Scanned check images.
– Output $\mathcal{Y}$: Positive numbers with two decimals.

Training a model

– Can we train a model using examples $(x_1, y_1) \ldots (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$?
– Possibly (we did not really try.)
– This would require excessive numbers of labeled examples.
– This would require excessive computation time.

engineering machine learning

Identify subproblems

- Locate amount fields.
- Segment characters in amount fields.
- Recognize isolated characters.
- Translate character strings into an amount.

Define a submodel for each subproblem

- Fairly complicated recognition models with large parameter vectors.
- Highly engineered location and segmentation models with only a few adjustable thresholds.

Collect and label data for each subproblem

- Lots of manual work.
- Manual labor is not very expensive. . .

engineering machine learning

Training strategies

– Independent training
  Train each submodel separately.
– Sequential training (better)
  Label outputs of submodel n and train submodel $n+1$.
– Global training (even better)
  Pre-train with sequential training.
  Simultaneously train all submodels with examples from $\mathcal{X} \times \mathcal{Y}$.

Issues

– The structure of the global model changes with the data.
  e.g. managing field location and segmentation hypotheses.
– The composition of submodels has nontrivial aspects.
  e.g. the label-bias problem.

deep learning

Deep learning (simplified)

– Pre-train with sequential unsupervised training
  Collect outputs of submodel n
  Train submodel $n+1$ with unsupervised criterion
– Tune with global training
  Consider all submodels as a single statistical model.
  Train with examples from $\mathcal{X} \times \mathcal{Y}$.

The deep learning surprise:

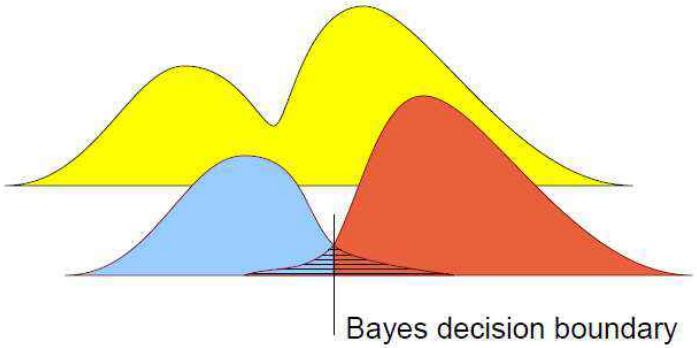
– Generic unsupervised subtasks work remarkably well.
– Little need to define subtasks using engineering knowledge.
– Little need to collect labeled data for all submodels.

Engineering learning systems is easier than we thought!

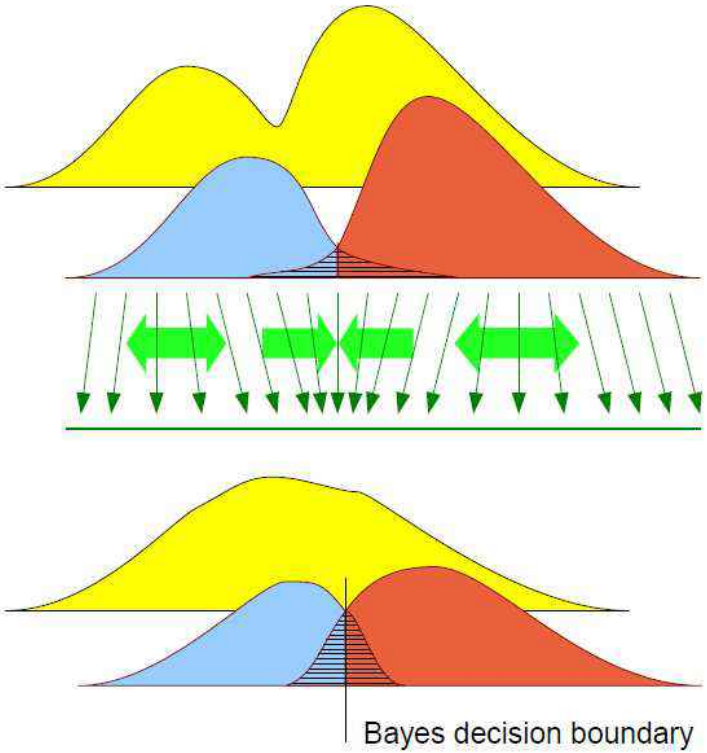
unsupervised learning

What is a cluster?

– Assumption: the shape of the density reveals the underlying categories.



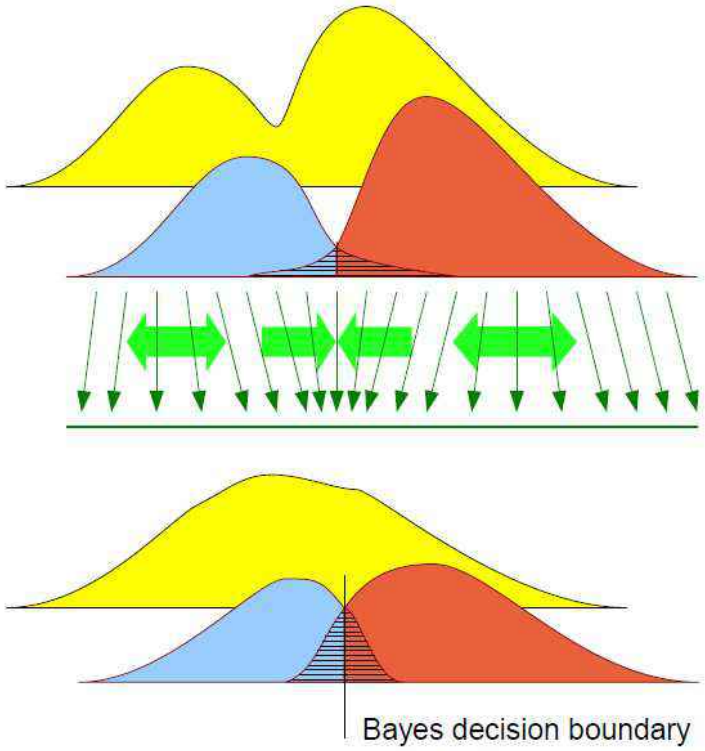
unsupervised learning



Input space transforms

- Categories are invariant.
- Bayes rate is invariant.
- Clustering is not invariant.

unsupervised learning



Clustering revisited

– Clustering is the expression of the prior knowledge encoded by our choice of input representation.

Unsupervised learning

– Comparable to using really cheap labels:
"x_1 and x_2 are close".
"x_1 and x_3 are not close".

auxiliary tasks

The price of labels

Interesting task $\iff$ Scarce labeled examples.
Uninteresting task $\iff$ Abundant labeled examples.

Auxilliary tasks

– **"In the vicinity of an interesting task** (with expensive labels,)
there often are less interesting tasks (with cheap labels,)
that we can put to good use."

– Unsupervised learning is just one of them (with trivial labels.)

– **Deep-learning, semi-supervised learning, and transfer learning are three facets of the same thing.**

e.g. (Weston et al., 2008)

example – face recognition

Interesting problem

– Recognizing the faces of one million persons.
– How many labeled images per person will we get?

Related but less interesting problem

– Are two face images representing the same person?
– Abundant (but noisy) examples:
  ◇ Two faces in the same image are likely to be different persons.
  ◇ Faces in successive frames are likely to be the same person.

(Matt Miller, NECLA, 2006)

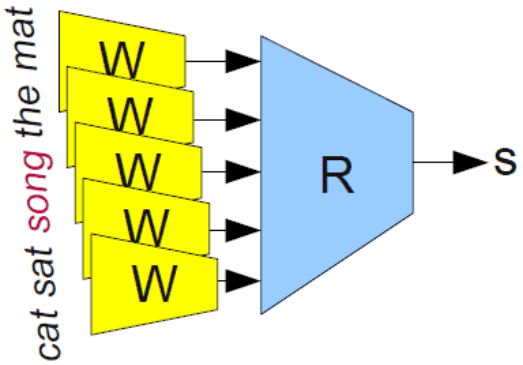
example – natural language tagging

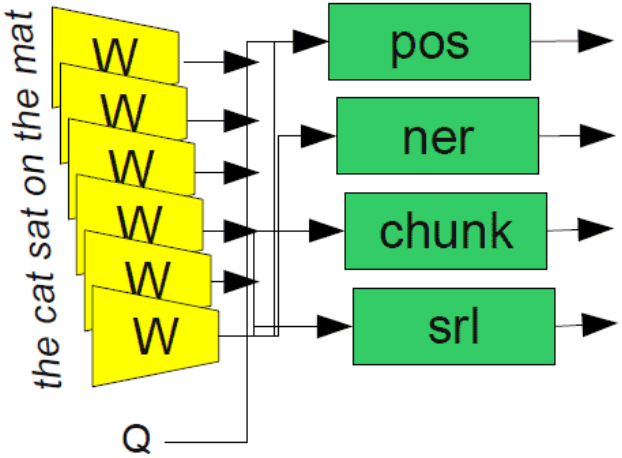
Interesting problems

– Standard NLP tagging tasks.

Related but less interesting problem

– Positive examples are legal sentence segments.
– Negative examples are created by substituting the central word.
– Ranking loss.





(Collobert et al., 2008-2011)

revisiting Vapnik's razor

" *When solving a (learning) problem of interest, do not solve a more complex problem as an intermediate step.* "

Rationale: how complex a model can we afford with our data?

However, solving a more complex task and transferring features often allows us to leverage more data of a different nature.

- Lots of implications.

  "From machine learning to machine reasoning", L.B., 2011.

conclusion

A good time to be in machine learning research.