Chapter6_应用与算法

1. 近似

1.1 范数近似

1.1.1 基本的范数逼近问题

- 基本形式（无约束问题）

  1.1 Norm Approximation

  $$\text{minimize} \quad \|Ax-b\|$$

  ($A \in \mathbf{R}^{m\times n}$ with $m \ge n$)

- 解释
  - 总的解释：回归问题
  - 几何解释：投影
  - 估计的解释：测量误差（越小越好）
  - 最优设计的解释
- 例示
  - 最小二乘问题（2-范数的平方） — 正规方程

    Least-squares approximation ($\|\cdot\|_2$): solution satisfies normal equations

    $$A^TAx = A^Tb$$

    ($x^\star = (A^TA)^{-1}A^Tb$ if $\mathbf{rank}\ A = n$)
  - Chebyshev逼近（∞-范数） — 描述为LP

    Chebyshev approximation ($\|\cdot\|_\infty$): can be solved as an LP

    $$\begin{array}{ll}\text{minimize} & t\\ \text{subject to} & -t\mathbf{1} \preceq Ax-b \preceq t\mathbf{1}\end{array}$$
  - 残差绝对值之和逼近（1-范数） — 描述为LP

    Sum of absolute residuals approximation ($\|\cdot\|_1$): can be solved as an LP

    $$\begin{array}{ll}\text{minimize} & \mathbf{1}^Ty\\ \text{subject to} & -y \preceq Ax-b \preceq y\end{array}$$

1.2 罚函数近似

- 基本形式

  1.2 Penalty function approximation.

  $$\begin{array}{ll}\text{minimize} & \phi(r_1)+\cdots\phi(r_m)\\ \text{subject to} & r = Ax-b\end{array}$$

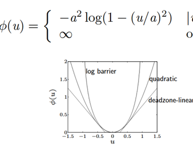
  ($A \in \mathbf{R}^{m\times n}$, $\phi : \mathbf{R}\to\mathbf{R}$ is a convex penalty function)

- **解释**：选定x之后，我们得到了b的一个逼近Ax，也得到了相应的残差向量r。罚函数评价残差每一个分量的惩罚。总体惩罚是每个分量残差惩罚之和。在罚函数逼近问题中，我们极小化残差带来的总体惩罚。

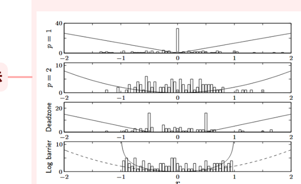
- **常见罚函数**

  Examples
  1. quadratic: $\phi(u) = u^2$
  2. deadzone-linear with width a: $\phi(u) = \max\{0, |u|-a\}$
  3. log-barrier with limit a:

  $$\phi(u) = \begin{cases} -a^2\log(1-(u/a)^2) & |u|<a \\ \infty & \text{otherwise}\end{cases}$$

  - 1. p-范数罚函数，p-范数逼近问题的等价形式，没有最外面的1/p次方
  - 2. 带有死区的线性罚函数，对于小于a的残差不进行惩罚
  - 3. 对数障碍罚函数，对于大于a的残差给予无穷的惩罚
  - 例示：l1-范数罚函数、l2-范数罚函数、死区线性罚函数和对数障碍罚函数的最优残差分布

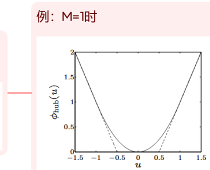
    

  - 4. Huber罚函数，又称鲁棒最小二乘，残差小于M时，与最小二乘罚函数相同；大于M时，恢复为线性增长

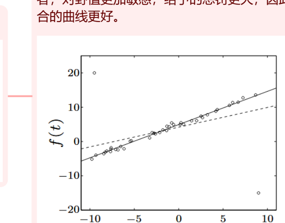
    Huber penalty function (with parameter M)

    $$\phi_{\text{hub}}(u) = \begin{cases} u^2 & |u|\le M \\ M(2|u|-M) & |u|>M\end{cases}$$

    - 例：M=1时

    - **Huber罚函数相对于l2-范数罚函数的效果**：前者为实线，后者为虚线，可以看到前者相对于后者，对野值更加敏感，给予的惩罚更大，因此拟合的曲线更好。

1.3 正则化近似

- **1. 双准则式**（w.r.t.是"关于"的意思）这两个范数可能是不同的：前者用来度量残差的规模；后者用来度量x的规模。目标是通过某些**办法**找到这两个目标之间的最优权衡。一个常用的办法就是**正则化。**

  1.3 Regularized approximation

  $$\text{minimize (w.r.t. } \mathbf{R}^2_+) \quad (\|Ax-b\|, \|x\|)$$

  ($A \in \mathbf{R}^{m\times n}$)

  interpretation: find good approximation $Ax \approx b$ with small x

  - **基本形式**（**两种方法**：引入问题参数或者极小化加权范数平方和）

    $$\text{minimize} \quad \|Ax-b\| + \gamma\|x\|$$

    Solution for $\gamma > 0$ traces out optimal trade-off curve

    other common method: minimize $\|Ax-b\|^2+\delta\|x\|^2$ with $\delta>0$

    - **Tikhonov正则化**：基于**极小化加权范数平方和**方法，选用l2-范数。这将得到一个（凸的）二次优化问题，且这个问题有解析解。

      Tikhonov regularization

      $$\text{minimize} \quad \|Ax-b\|_2^2 + \delta\|x\|_2^2$$

      can be solved as a least-squares problem

      $$\text{minimize} \quad \left\| \begin{bmatrix} A \\ \sqrt{\delta}I\end{bmatrix} x - \begin{bmatrix} b \\ 0\end{bmatrix}\right\|_2^2$$

      solution $x^\star = (A^TA+\delta I)^{-1}A^Tb$

- **2. 应用实例**：最优输入设计

  考虑一个动态系统，u(·)是输入标量序列，y(·)是输出标量序列，它们通过卷积相互关联（第一个式子）
  序列h(·)称为卷积核或者系统的脉冲响应。
  目标：1.跟踪输出（输出y需要逼近给定目标或者参考信号y_{des}，用J_{track}表示）
  2.输入尽量小（用J_{mag}表示）
  3.输入的变化尽量小（用J_{der}表示）
  通过最小化加权和我们可以在目标间进行权衡

  linear dynamical system with impulse response h:

  $$y(t) = \sum_{\tau=0}^{t} h(\tau)u(t-\tau), \quad t = 0,1,\ldots,N$$

  Input design problem: multicriterion problem with 3 objectives
  1. tracking error with desired output y_{des}: $J_{\text{track}} = \sum_{t=0}^{N}(y(t)-y_{\text{des}}(t))^2$
  2. input magnitude: $J_{\text{mag}} = \sum_{t=0}^{N} u(t)^2$
  3. input variation: $J_{\text{der}} = \sum_{t=0}^{N-1}(u(t+1)-u(t))2$

  track desired output using a small and slowly varying input signal

  regularized least-squares formulation

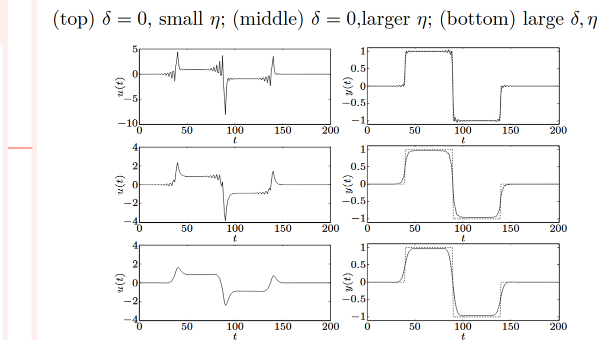
  $$\text{minimize} \quad J_{\text{track}} + \delta J_{\text{der}} + \eta J_{\text{mag}}$$

  for fixed δ, η, a least-squares problem in $\mu(0),\ldots,\mu(N)$

  - 考虑一个具体的例子：

    example: 3 solutions on optimal trade-off curve

    (top) $\delta = 0$, small η; (middle) $\delta = 0$,larger η; (bottom) large δ,η

- **3. 信号重建**

  基本形式（其中w.r.t译为关于）：原始信号x在被未知、小值且快速变动的噪声v干扰之后，得到信号x_{cor}。目标是在给定受污染信号x_{cor}的情况下，构建对原始信号x的估计值hat{x}。重构问题的一个简单方式是如下图所示的双准则问题，其中第二项称为正则化函数或光滑目标，用来评估hat{x}的粗糙程度。

  1.3.3 Signal reconstruction

  $$\text{minimize (w.r.t. } \mathbf{R}^2_+) \quad (\|\hat{x}-x_{\text{cor}}\|_2, \phi(\hat{x}))$$

  $x \in \mathbf{R}^n$ is unknown signal

  $x_{\text{cor}} = x+v$ is (known) corrupted version of x, with additive noise v

  variable $\hat{x}$ (reconstructed signal) is estimate of x

  $\phi : \mathbf{R}^n \to \mathbf{R}$ is regularization function or smoothing objective

  - **构建双准则问题的方法：**
    1. 二次光滑
    2. 总变差重构

    二次光滑

    最简单的重构方法是使用二次光滑函数

    $$\phi_{\text{quad}}(x) = \sum_{i=1}^{n-1}(x_{i+1}-x_i)^2 = \|Dx\|_2^2,$$

    其中，$D \in \mathbf{R}^{(n-1)\times n}$ 为双对角矩阵

    $$D = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{bmatrix}.$$

    我们可以通过极小化

    $$\|\hat{x}-x_{\text{cor}}\|_2^2 + \delta\|D\hat{x}\|_2^2$$

    得到 $\|\hat{x}-x_{\text{cor}}\|_2$ 和 $\|D\hat{x}\|_2$ 之间的最优权衡，其中 $\delta>0$ 参数化了最优权衡曲线。这个二次问题的解

    $$\hat{x} = (I+\delta D^TD)^{-1}x_{\text{cor}}$$

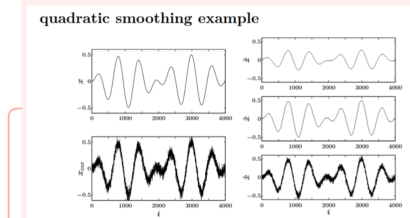
    可以非常高效地求得，因为 $I+\delta D^TD$ 是三对角的；参见附录 C。

    examples: quadratic smoothing, total variation smoothing:

    $$\phi_{\text{quad}}(\hat{x}) = \sum_{i=1}^{n-1}(\hat{x}_{i+1}-\hat{x}_i)^2, \quad \phi_{\text{tv}}(\hat{x}) = \sum_{i=1}^{n-1}|\hat{x}_{i+1}-\hat{x}_i|$$

    - **二次光滑应用实例**
      原始信号光滑且噪声变化很快时，二次光滑是良好的方法

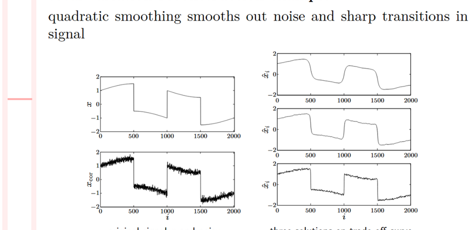
      quadratic smoothing example

      - **二次光滑的缺陷**
        原始信号的任意快速变化都会被该方法减弱或者移除。

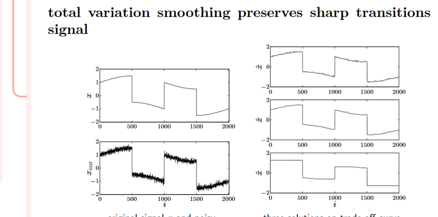
        total variation reconstruction example

        quadratic smoothing smooths out noise and sharp transitions in signal

    - **总变差重构的例子**
      不像二次光滑，总变差重构中信号的尖锐变化被保留了下来

      total variation smoothing preserves sharp transitions in signal

1.4 鲁棒近似

- **基本形式**：在矩阵A未知的情况下优化目标函数‖ Ax - b ‖
- **两种思路**
  1. Stochastic: assume A is random, minimize $\mathbf{E}\|Ax-b\|$
  2. Worst-case: set $\mathcal{A}$ of possible values of A, minimize $\sup_{A\in\mathcal{A}}\|Ax-b\|$

  tractable only in special cases (certain norms $\|\cdot\|$, distributions, sets $\mathcal{A}$)

  - **随机鲁棒逼近**：优化目标函数的期望及其等价形式（比如期望的平方，比较容易处理，也叫随机鲁棒最小二乘问题）
    - **随机鲁棒最小二乘**
      图片给出了问题的目标函数、问题的解析解以及其与Tikhonov问题的等价性证明

      Stochastic robust LS with $A = \bar{A}+U$, U random, $\mathbf{E}U = 0$, $\mathbf{E}U^TU = P$

      $$\text{minimize} \quad \mathbf{E}\|(\bar{A}+U)x-b\|_2^2$$

      1. explicit expression for objective:

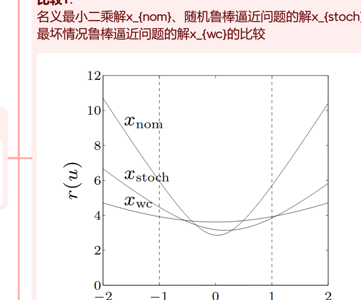
      $$\begin{aligned}\mathbf{E}\|Ax-b\|_2^2 &= \mathbf{E}\|\bar{A}x-b+Ux\|_2^2\\ &= \|\bar{A}x-b\|_2^2 + \mathbf{E}x^TU^TUx\\ &= \|\bar{A}x-b\|_2^2 + x^TPx\end{aligned}$$

      2. hence, robust LS problem is equivalent to LS problem

      $$\text{minimize} \quad \|\bar{A}x-b\|_2^2 + \|P^{1/2}x\|_2^2$$

      3. for $P = \delta I$, get Tikhonov regularized problem

      $$\text{minimize} \quad \|\bar{A}x-b\|_2^2 + \delta\|x\|_2^2$$
  - **最坏情况鲁棒逼近**：优化目标为 sup{‖Ax-b‖ | 对所有可行的A}
  - **比较1**：
    名义最小二乘解x_{nom}、随机鲁棒逼近问题的解x_{stoch}和最坏情况鲁棒逼近问题的解x_{wc}的比较

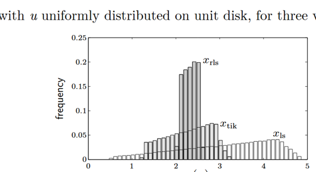
  - 比较2：
    最坏情况鲁棒问题的解、Tikhonov正则化问题的解和名义最小二乘解的比较

    名义最小二乘解残差分布广泛，且最小二乘解对于参数的变化非常敏感；鲁棒最小二乘和Tikhonov正则化的解展现出了对于整个单位元上参数的不确定性的相当小的变化。

    example

    $$r(u) = \|(A_0+u_1A_1+u_2A_2)x-b\|2$$

    with u uniformly distributed on unit disk, for three values of x

    1. x_{ls} minimizes $\|A_0x-b\|_2$
    2. x_{tik} minimizes $\|A_0x-b\|_2^2+\delta\|x\|_2^2$ (Tikhonov solution)
    3. x_{wc} minimizes $\sup_{\|u\|_2\le 1}\|A_0x-b\|_2^2+\|x\|_2^2$

2. 统计估计

2.0 前置知识

- **分布估计问题**：根据观测值估计随机变量的概率密度函数
- **参数估计问题**：choose from a family of densities p_x(y), indexed by a parameter x

2.1 最大似然估计

- **基本思想**：最大化对数似然函数

  $$\text{maximize (over } x) \quad \log p_x(y)$$

  1. y is observed value
  2. $l(x) = \log p_x(y)$ is called log-likelihood function
  3. can add constraints $x \in C$ explicitly, or define $p_x(y) = 0$ for $x \notin C$
  4. a convex optimization problem if $\log p_x(y)$ is concave in x for fixed y

- **附加了IID（独立同分布）噪声的线性测量模型**
  x是带估计参数向量，y_i是测量量或者观测量，v_i是测量误差或者噪声（设为独立同分布），在R上具有概率p。似然函数如下，可以找到如下最优点。如果概率密度p是对数-凹的，此优化问题是凸优化问题，且与罚函数逼近问题形式一样（此时罚函数为-log p）。

  Linear Measurement Model

  $$y_i = a_i^Tx + v_i, \quad i = 1,\ldots,m$$

  $x \in \mathbf{R}^n$ is vector of unknown parameters

  v_i is IID measurement noise, with density $p(z)$

  y_i is measurement: y has density $p_x(y) = \Pi_{i=1}^m p(y_i - a_i^Tx)$

  maximum likelihood estimate: any solution x of

  $$\text{maximize (over } x) \quad l(x) = \textstyle\sum_{i=1}^m \log p(y_i - a_i^Tx)$$

  y is observed value

- **噪声服从一些常见概率分布时的ML估计**
  1.高斯分布（正态分布）
  2.Laplace分布
  3.均匀分布

  2 Statistical Estimation

  1. **Gaussian noise** $\mathcal{N}(0,\sigma^2)$: $p(z) = (2\pi\sigma^2)^{-1/2}e^{-z^2/(2\sigma^2)}$,

  $$l(x) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(a_i^Tx-y_i)^2$$

  ML estimate is LS solution

  2. **Laplacian noise**: $p(z) = (1/(2a))e^{-|z|/a}$,

  $$l(x) = -m\log(2a) - \frac{1}{a}\sum_{i=1}^m |a_i^Tx - y_i|$$

  ML estimate is 1-norm solution

  3. **uniform noise on** $[-a, a]$:

  $$l(x) = \begin{cases} -m\log(2a) & |a_i^Tx-y_i|\le a, \ i = 1,\ldots,m \\ -\infty & \text{otherwise}\end{cases}$$

  ML estimate is any x with $|a_i^Tx - y_i| \le a$

- **Logistic回归**

  Logistic regression

  Random variable $y \in \{0,1\}$ with distribution

  $$p = \mathbf{prob}(y=1) = \frac{\exp(a^Tu+b)}{1+\exp(a^Tu+b)}$$

  1. a, b are parameters; $u \in \mathbf{R}^n$ are (observable) explanatory variables
  2. Estimation problem: estimate a, b from m observations (u_i, y_i)

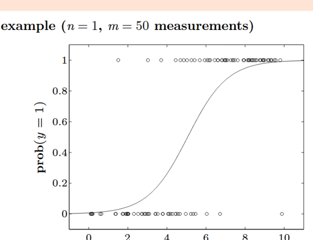
  log-likelihood function (for $y_1 = \cdots = y_k = 1$, $y_{k+1} = \cdots = y_m = 0$):

  $$\begin{aligned} l(a,b) &= \log\left(\Pi_{i=1}^k \frac{\exp(a^Tu_i+b)}{1+\exp(a^Tu_i+b)} \Pi_{i=k+1}^m \frac{1}{1+\exp(a^Tu_i+b)}\right) \\ &= \sum_{i=1}^k (a^Tu_i+b) - \sum_{i=1}^m \log(1+\exp(a^Tu_i+b))\end{aligned}$$

  concave in a, b

  - eg.

    example ($n=1$, $m=50$ **measurements)**

    circles show 50 points (u_i, y_i)

    solid curve is ML estimate of $p = \exp(au+b)/(1+\exp(au+b))$

*2.2 最优探测器设计

3. 算法：无约束优化问题

3.1 Decent

- **无约束优化的算法描述**
  1. 决定优化方向
  2. 决定优化步长（这一步也叫**直线搜索**）
  3. 修改x的值

  $$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)} \quad \text{with} \quad f(x^{(k+1)}) < f(x^{(k)})$$

  1. other notations: $x^+ = x + t\Delta x$, $x := x + t\Delta x$
  2. Δx is the step, or search direction; t is the step size, or step length
  3. from convexity, $f(x^+) < f(x)$ implies $\nabla f(x)^T\Delta x < 0$ (i.e., Δx is a descent direction)

- **直线搜索方法**
  - **精确直线搜索**：找δx方向上使得函数值最小的t的值，即为步长

    $$t = \arg\min_{t>0} f(x+t\delta x)$$
  - **回溯直线搜索**：从单位步长开始，按比例逐渐减小，直到满足停止条件。

- **确定优化方向的方法：梯度下降方法**
  即使用负梯度作为优化方向

  - **given** a starting point $x \in \textbf{dom}\ f$.
    repeat
    1. $\Delta x := -\nabla f(x)$.
    2. *Line search.* Choose step size t via exact or backtracking line search.
    3. *Update.* $x := x + t\Delta x$.

    until stopping criterion is satisfied.

    上为梯度下降算法描述
    1. 停止条件通常为梯度足够小
    2. 很简单，但是很慢，实际工作中不太用它
    3. 结果收敛

  - **例子**
    - **R^2空间的二次问题**

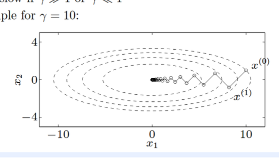
      $$f(x) = (1/2)(x_1^2+\gamma x_2^2) \quad (\gamma>0)$$

      with exact line search, starting at $x^{(0)} = (\gamma, 1)$:

      $$x_1^{(k)} = \gamma\left(\frac{\gamma-1}{\gamma+1}\right)^k, \quad x_2^{(k)} = \left(-\frac{\gamma-1}{\gamma+1}\right)^k$$

      very slow if $\gamma \gg 1$ or $\gamma \ll 1$

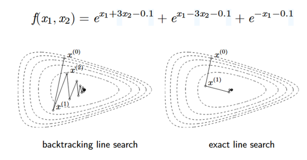
      example for $\gamma = 10$:

    - **R^2空间的非二次问题**

      nonquadratic example

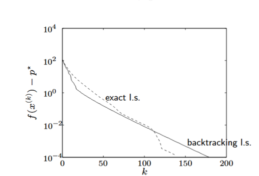
      $$f(x_1,x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$$

    - **R^100空间的一个问题**

      a problem in $\mathbf{R}^{100}$

      $$f(x) = c^Tx - \sum_{i=1}^{500}\log(b_i - a_i^Tx)$$

      'linear' convergence, i.e., a straight line on a semilog plot

- **最速下降方法**
  - **规范化最速下降方法**

    normalized steepest descent direction (at x, for norm $\|\cdot\|$):

    $$\Delta x_{\text{nsd}} = \arg\min\{\nabla f(x)^Tv \mid \|v\| = 1\}$$

    interpretation: for small v, $f(x+v) \approx f(x) + \nabla f(x)^Tv$

    direction Δx_{nsd} is unit-norm step with most negative directional derivative
  - **非规范化最速下降方法**

    (unnormalized) steepest descent direction

    $$\Delta x_{\text{sd}} = \|\nabla f(x)^T\|_* \Delta x_{\text{nsd}}$$

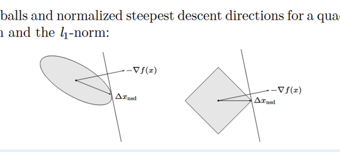
    satisfies $\nabla f(x)^T\Delta x_{\text{sd}} = -\|\nabla f(x)^T\|_*^2$
  - **最速下降模型**

    steepest descent method
    1. general descent method with $\Delta x = \Delta x_{\text{sd}}$
    2. convergence properties similar to gradient descent

    - **例子**

      Examples
      1. Euclidean norm: $\Delta x_{\text{sd}} = -\nabla f(x)$
      2. Quadratic norm $\|x\|_P = (x^TPx)^{1/2} (P \in \mathbf{S}^n_{++})$: $\Delta x_{\text{sd}} = -P^{-1}\nabla f(x)$
      3. l_1-norm: $\Delta x_{\text{sd}} = -(\partial f(x)/\partial x_i)e_i$, where $|\partial f(x)/\partial x_i| = \|\nabla f(x)\|_\infty$

      unit balls and normalized steepest descent directions for a quadratic norm and the l_1-norm:

3.2 牛顿法

3.2 Newton step

$$\Delta x_{\text{nt}} = -\nabla^2 f(x)^{-1}\nabla f(x)$$

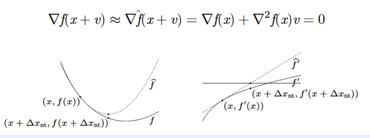
interpretations

1. $x + \Delta x_{\text{nt}}$ minimizes second order approximation

$$\hat{f}(x+v) = f(x) + \nabla f(x)^Tv + \frac{1}{2}v^T\nabla^2 f(x)v$$

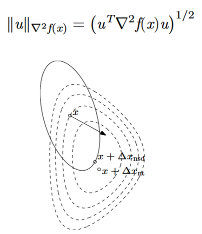
2. $x + \Delta x_{\text{nt}}$ solves linearized optimality condition

$$\nabla f(x+v) \approx \nabla\hat{f}(x+v) = \nabla f(x) + \nabla^2 f(x)v = 0$$



- Δx_{nt} is steepest descent direction at x in local Hessian norm

  $$\|u\|_{\nabla^2 f(x)} = \left(u^T\nabla^2 f(x)u\right)^{1/2}$$

  dashed lines are contour lines of f; ellipse is $\{x + v \mid v^T\nabla^2 f(x)v = 1\}$

  arrow shows $-\nabla f(x)$